Analyse des Corrélations entre Pannes dans les Systèmes de Stockage Pair-à-Pair †

O. Dalle and F. Giroire and J. Monteiro and S. Pérennes

MASCOTTE joint project team, INRIA, I3S, CNRS, Univ. Nice Sophia - B.P. 93, F-06902 Sophia Antipolis, France.

Dans cet article, nous présentons et étudions des modèles analytiques de systèmes de stockage pair-à-pair fiables à long terme. Les pairs sont sujets à des pannes définitives (défaillance du disque, départ du pair) induisant la perte de toutes les données stockées par le pair. Ces pannes ont lieu en continu. Afin de pérenniser les données il est indispensable d'user de redondance et de maintenir celle-ci au moyen d'un processus permanent de *reconstruction*. Dans un premier temps nous considérons une approche classiquement utilisée dans la littérature, consistant à modéliser chaque bloc par une chaîne de Markov et à négliger les interdépendances entre blocs. Si celle-ci permet le calcul du comportement moyen du système (par exemple la demande moyenne en bande passante), elle est insuffisante pour en évaluer les fluctuations. Nos simulations démontrent que ces fluctuations sont très importantes même pour des grands systèmes comportant des milliers de pairs. Nous proposons alors un nouveau modèle stochastique prenant en compte l'interdépendance des pannes de blocs, et nous en donnons une approximation fluide. Ceci nous permet de caractériser le comportement du système (calcul de tous les moments) mais aussi de le simuler efficacement, car il est indépendant de la taille du système. La pertinence de notre modèle est validée en comparant les résultats obtenus par des simulations utilisant d'un côté notre modèle fluide et de l'autre un modèle à événements discrets reproduisant fidèlement le comportement du système.

Keywords: P2P storage system, failure correlation, performance, data durability, Markov chain model, fluid model

1 Introduction

In this paper‡, we study peer-to-peer storage systems that have high durability requirements (i.e., backup systems or long-term storage systems), like Intermemory, CFS, Farsite, OceanStore, PAST, Total-Recall. To achieve high resilience over a long period of time, such P2P systems encode the user data in a set of redundant fragments and distribute those fragments among the peers. We consider here systems using Erasure Codes for redundancy, as they usually have a lower storage space overhead than replication [3].

We study the following questions: How much resource (bandwidth and storage space) is necessary to maintain redundancy and to ensure a given level of reliability? What is the probability that the system loses data? To address those questions, we first define a Markov Chain Model (MCM) that represents the behavior of a single data block. This chain allows to compute the average behavior of the system accurately.

Simulations confirm our analytical results, but also indicate that the variations around the average behavior are much higher than in the MCM. These variations are explained by the fact that when a disk failure occurs, many data fragments are lost *at the same time*. This induces large peaks in the bandwidth demand. In addition, when the bandwidth is limited, those peaks tend to slow down the reconstruction process, which results in data losses. Indeed, when the reconstruction time is longer, a damaged block is more likely to lose its remaining redundancy fragments. The consequence is that a bandwidth provisioning decision not taking into account these variations would lead to a significant loss of data.

In order to take into account this phenomenon, we propose a new stochastic Approximated Model, that does not represent a single block anymore, but the whole system. We provide a mathematical analysis

†This work was partially funded by the European project IST FET AEOLUS and the ANR PROJECT SPREADS.

‡ An extended version can be found in the INRIA Research Report RR-6771, `http://hal.inria.fr/inria-00346857/`

of this model by giving a method to compute all the moments of its associated stationary distribution. Simulations show that the Approximated Model predicts the system very well (1% margin). Moreover, this Approximated Model is scalable since its complexity is proportional to the erasure code length and does not depend on the number of peers. Last, we present a fluid approximation that reduces the study of this model to the random product of two simple matrices.

Related Work: The literature about P2P storage systems is abundant and several systems have been proposed. However, few analytical models have been studied to estimate accurately the behavior of those systems (data durability, resource usage, e.g., bandwidth) and understand the trade-offs between the system parameters. DHT based systems have been studied formally (see Karger et al., 2002) but they have different requirements (e.g. network connectivity instead of data durability) and they assume replication is used for redundancy. In [3], the authors show that, in most cases, erasure codes use an order of magnitude less bandwidth and storage than replication to provide similar system durability. In [1], the authors use a Markovian analysis to evaluate the performance of systems using Erasure codes for two different schemes of data recovery (centralized vs. distributed) and estimate the data lifetime and availability. In all these models, block failures are considered independent.

System Description: We consider a system designed for data archival. In this case the user data is immutable and stays for ever in the system. Furthermore, we study the steady state with constant number of peers. A peer can leave the system for short periods of time (short-time churn). Peers are subject to failures, mainly disk crashes. They get *faulty independently* according to a memoryless Poisson process. When a disk fails, it is replaced with a new empty one.

The data is divided into user data blocks. Each user data block is, in turn, sub-divided into s equally sized *fragments* to which are added r fragments of redundancy, using Erasure Codes. Each block has then $n = s + r$ fragments that are spread and stored on n different peers chosen at random. *Any* subset of s fragments chosen among the $s + r$ initial fragments is sufficient to recover (reconstruct) the block.

When a block b has less than r_0 fragments of redundancy left, its reconstruction is initiated (in the results shown here $s = 9$, $r = 6$ and $r_0 = 3$). A peer is then chosen uniformly at random to carry out the reconstruction. First, it downloads s of the remaining fragments, then it rebuilds the block, and finally it spreads the missing fragments in the network. Monitoring the state of the system is needed to decide which blocks have lost a critical number of their fragments. This can be done with a Distributed Hash Table (DHT) where peers are responsible for a subset of other peers.

2 Limitations of Markov Chain Models

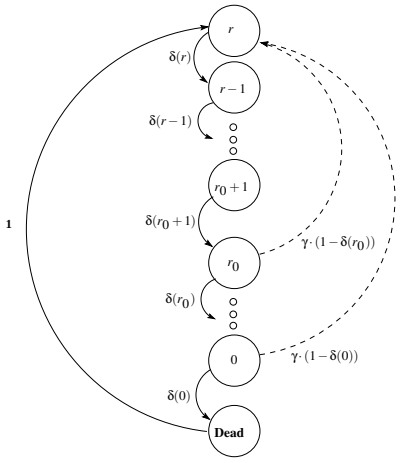

Figure 1: Markov chain modeling the behavior of one block. Solid and dashed lines represent respectively failures and reconstructions events.

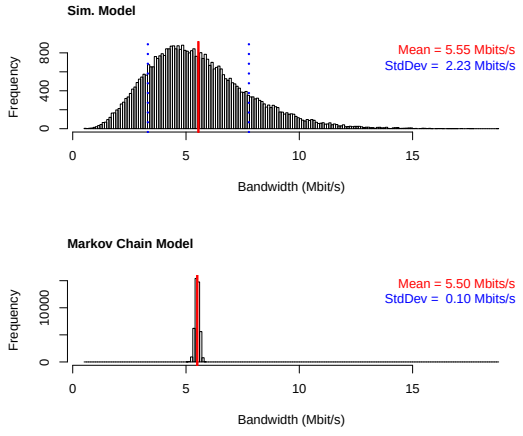

Figure 2: Histogram of the bandwidth used by the system. Top: System with disk failures. Bottom: system with independent block failures.

Markov Chain Model (MCM). We propose an MCM for our specific peer-to-peer storage system, similar to those found in the literature [1]. The chain (as depicted in Figure 1) has $r+2$ states. $r+1$ states represent a

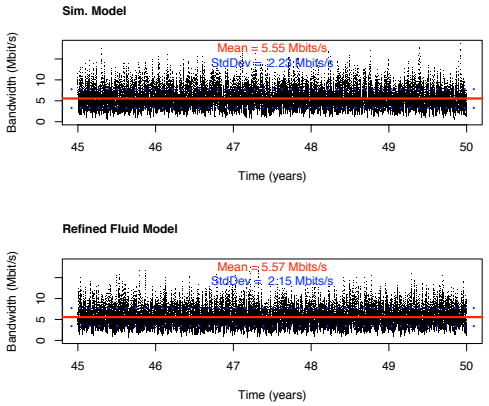

Figure 3: Time series of the bandwidth used by SM and RFM for 5 years. The behavior of both models are very close.

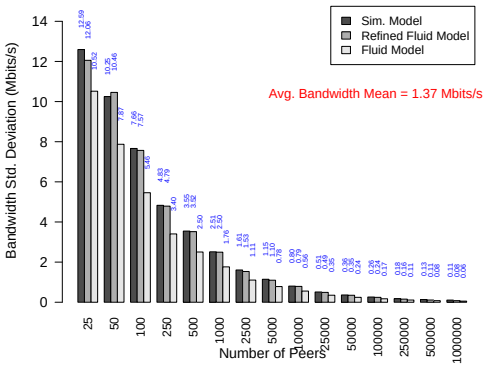

Figure 4: The std. deviation of bandwidth usage is roughly the same for the SM and RFM (we see the improvement from FM). Note that the values for MCM (0.07, not shown) are constant for all values.

level of redundancy of the block, $r(b)$, and the last one represents a *Dead* state. The probability for a block at level i to lose one fragment during a time step is denoted by $\delta(i)$ and is given by $\delta(i) := (s+i)f(1-f)^{s+i-1}$, where f is the probability for a disk to experience a failure during a time step. When a block becomes critical ($r(b) \leq r_0$), the reconstruction starts. The reconstruction is modeled as follows: the average duration of a reconstruction being noted θ, at each time step, a critical block has a probability $\gamma := 1/\theta$ to be rebuilt, and in that case it goes to the top. If a block loses more than the available redundancy fragments before being reconstructed, it goes to the *dead* state. This finite Markov chain is irreducible and aperiodic. Hence, it admits a unique stationary distribution that can be computed exactly in time polynomial in n.

Results. We performed a large number of simulations with different sets of parameters on a custom cycle-based simulator. The results shown here are for a system with 5000 peers, each containing 1500 fragments. The simulation time is 10 years with a time step of one hour. Figure 2 gives the experimental distribution of the bandwidth usage in two cases: the simulation of the real system (top plot) and the simulation of a system with independent block failures, as modeled by the MCM (bottom plot). The average value of both systems are very close (5.55 versus 5.50 Mbits/s). However, the variations around this average are totally different. The standard deviation is 2.23 Mbits/s in the first case, compared to only 0.1 Mbits/s in the second case. Hence, we see that the impact of failure correlation is very strong on the behavior of the system. Additionally, a bandwidth provisioning decision that does not take into account these variations leads to a significant loss of data[‡].

3 A New Stochastic Fluid Model

The discussion above shows that the systems cannot be seen as a set of independent blocks; so we need to model the system globally. However, using again the same approach would result in a gigantic Markov Chain, in order to reflect the location of all fragments (about $N^{B\cdot(s+r)}$ states). Therefore, this chain cannot be used in practice to implement a simulator. We propose here a Approximated Model whose purpose is too approximate this gigantic chain. We also present a theoretical analysis that allows to compute all the moments of its stationary distribution. Then, we introduce two fluid approximations (Fluid Model, Refined Fluid Model) of the Approximated Model. The analysis of the two fluid models boils down to the analysis of the random product of two simple matrices, $M(t,\omega)$. Note that we do not give a closed formal solution to this difficult problem because there exists no general theory to get the distribution of a random product of two matrices. It is not surprising since, for example, only determining if the infinite product of two matrices is null is an undecidable problem [2].

The Approximated Model. The Approximated Model is derived from the following observation: fragments are spread randomly at the system start-up time and whenever a reconstruction occurs. Hence, we make the following assumption:

(A) At any time the fragments of a block are randomly placed into the system[§].

In such a case, the state of a block is fully described by its level. We can then describe the system by a vector $B(t) = (B_0(t), \cdots, B_r(t))$ where $B_i(t)$ is the number of blocks at level i at time t. The system dynamics can then be described by a random product of matrices. The system is scalable since its size is $s+r$ and the random transition matrix at time t can be computed in time $(s+r)^2$. Due to lack of space we don't give it here explicitly, instead we present its fluid approximation which is simpler.

The Fluid Model. We describe the system by the vector $X(t) = (X_0(t), \cdots, X_r(t))$, where $X_i(t)$ counts the percentage of blocks that are in state i at discrete time t (i.e., $X(t) = B(t)/B$). First, we define the matrix F' (resp. R) which represents the effects of a disk failure (resp. of the reconstruction process) on the state vector

$$F' = \begin{pmatrix} 1-\mu(t,\omega) & & & & \mu(t,\omega) \\ \mu(t,\omega) & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \mu(t,\omega) & 1-\mu(t,\omega) \end{pmatrix} \quad R = \begin{pmatrix} 1 & & \gamma & \cdots & \gamma \\ & \ddots & & & \\ & & 1 & & \\ & & & 1-\gamma & \\ & & & & \ddots \\ & & & & 1-\gamma \end{pmatrix}$$

where $\mu_i(t,\omega)$ is the fraction of blocks in state i affected by a failure. We then express a transition of the system as $X(t+1,\omega) = M(t,\omega) \cdot X(t,\omega)$, with $M(t,\omega)$ a random product defined as follows

$$M(t,\omega) = \begin{cases} F'R & \text{with probability } l \text{ (disk failure);} \\ R & \text{with probability } 1-l \text{ (reconstruction only),} \end{cases}$$

where l is the probability to experience a disk failure during a time step. At each time step, if no disk failure occurs, we only account for the effects of the reconstructions; otherwise the disk failure effect is added.

In this model, we assume that, whenever there is a failure, a block at level i has probability $\mu_i(t,\omega)$ to lose a fragment. This is indeed hypothesis (A). A first approach is then to consider that each disk contains a proportion $1/N$ of fragments (i.e., about $B(s+r)/N$), then the probability to lose a fragment at level i (assuming a failure) is $\mu_i(t,\omega) \mid failure = \frac{s+i}{N}$. This approach already gives good results but we can refine it. Since disks fill up during the system life, a newly replaced disk is empty, while an old disk contains many fragments. Computing the disk age and disk size distributions is easy (geometric laws); so we can take it into account and modify $\mu_i(t,\omega)$ accordingly. This can be done by setting $\mu_i(t,\omega) = \frac{(s+i)z(\omega)}{N}$ where $z(\omega)$ is taken according to the distribution of the numbers of fragments in a disk. This yields the Refined Fluid Model.

Analysis Using the fact that $X(t+1,\omega) = M(t,\omega)X(t)$ we can compute all moments of the distribution of $X(t)$. As example $\mathbb{E}(X(t))$ converges to the unique eigenvector of $\mathbb{E}[M(t,\omega)] := lF + (1-l)R$. Since $(lF + (1-l)R)$ is equivalent to the matrix transition of the single block MCM we find that $\mathbb{E}(X(t))$ converges to the stationary vector of the single block model, this is natural since expectations are linear. Other moments can be computed similarly, albeit with additional complexity, as we need to compute all cross-products ($E[X_1 \cdot \ldots \cdot X_k]$ for the k-th moment).

Validation: Simulations validate the Fluid Model, see Figure 3 (system bandwidth usage) and Figure 4 (standard deviation for different numbers of peers).

References

[1] S. Alouf, A. Dandoush, and P. Nain. Performance analysis of peer-to-peer storage systems. *Internation Teletraffic Congress (ITC), LNCS 4516*, 4516:642–653, 2007.

[2] A. Markov. On the problem of representability of matrices. *Z. Math. Logik Grund. Math*, pages 157–168, 1958.

[3] H. Weatherspoon and J. Kubiatowicz. Erasure coding vs. replication: A quantitative comparison. In *Proc. of IPTPS*, volume 2, pages 328–338. Springer, 2002.

[§] Assumption (A) is indeed incorrect since the fragments of a block whose last reconstruction occurred at time T_0 can only be located on the disks that where in the system at time T_0 and never got faulty since. The correct statement is that the fragments of a block with age $T - T_0$ are randomly spread on disks with age at least $T - T_0$. Nevertheless we will assume that (A) holds.